Jan 21, 2008

Under the Hood: 45 nm: What Intel didn't tell you

Don Scansen, Semiconductor Insights

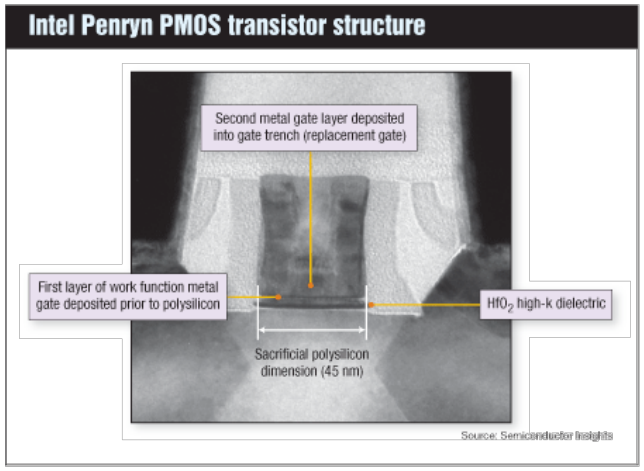
As noted in EE Times almost one month prior to the December 2007 International Electron Devices Meeting (IEDM), the main features of Intel's 45-nanometer process technology are the incorporation of high-k hafnium-based dielectric material, titanium nitride (TiN) for the PFET replacement gate and a TiN barrier alloyed with a work function tuning metal for the NFET replacement gate. (Click here for full story.)

Some high points of Intel's 45-nm HKMG technology are: high-k first, metal-gate-last integration; hafnium oxide (HfO_2) gate dielectric (1.0-nm EOT); and dual band-edge work function metal gates (TiN for PMOS; TiAlN for NMOS).

The gate-last integration is one point that needs a bit of clarification in the Intel process flow.

The references to "first" and "last" refer to the order of the high-k and metal-gate formation in light of polysilicon deposition. It is now well-known that Intel uses a gate-last, or replacement gate, process flow at 45 nm. But even that provides opportunity for a great debate on semantics: whether it's "gate" or "last." I'm not predicting that the lawyers are already on their way, but there's likely a patent out there that will help create just such an argument.

The replacement gate flow allows Intel to reuse many process steps and tools from the age-old polysilicon-gate technology. The process of patterning polysilicon and forming traditional silicon oxide and nitride sidewall spacers leverages tried-and-true self-aligned processes for source and drain formation and their lightly doped extension regions. Once those steps are completed, the polysilicon is removed, and work function metals are deposited in their stead.

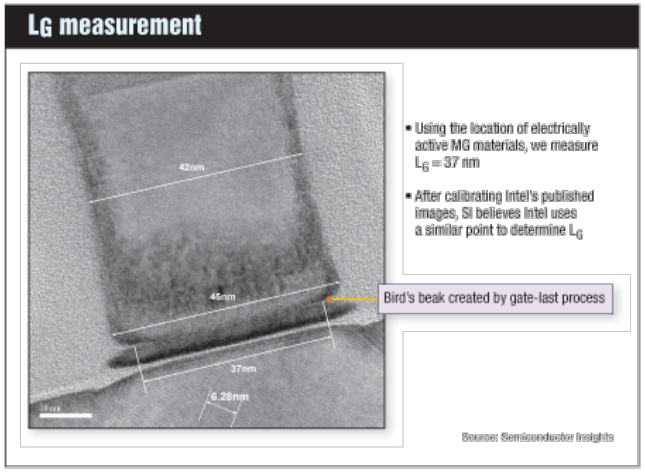


Click here for larger image

But something interesting is going on even before the first poly deposition. Contrary to the suggestion in its IEDM paper, Intel deposits the first work function metal prior to the sacrificial gate polysilicon.

For the P-channel transistor, titanium nitride (TiN) is deposited immediately after the HfO_2 dielectric. Adding aluminum to form TiAlN tunes the work function for

the N-channel transistors. Intel's process protects HfO_2 from the polysilicon etch by depositing the first work function layers before forming and patterning polysilicon. SI engineers refer to the first metal-gate layer as the top interface layer (TIL) because of the undeniable protection it provides the HfO_2 dielectric. The P-type metal gates are TiN, and Al is added to create TiAlN and the appropriate work function for NMOS. Thicker layers of both metals are deposited in their respective N- and P-channel transistors after removing the sacrificial polysilicon, and a barrier layer is formed on the bottom and sidewalls of the trench left behind by the polysilicon etch.



Click here for larger image

Comments about the meaning of "gate" are arguably less important than the electrical performance of the finished product. Intel's 45-nm technology is certainly impressive in that regard. SI's extraction of transistor electrical parameters indicates the following saturated drive currents at 1.0 V at room temperature:

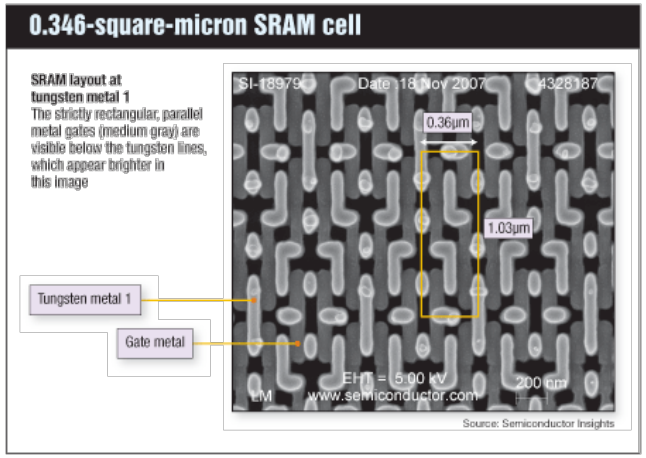
• PFET IDSAT = 1.08 mA/micrometer

• NFET IDSAT = 1.36 mA/µm

Intel confirmed these values at its IEDM presentation in December (although our PFET number is actually 10 µA higher than Intel reported). Not surprisingly, our results show higher drive currents at low temperature (–20°C) and reduced current at high temperature (85°C).

These high values for drive current evoke more questions regarding the gate structure. There has always been a discrepancy between the physical gate length (L_G) of transistors and the shorter electrically active channel length (L_{elec}). Before the advent of modern metal-gate technology, however, it was relatively easy to specify L_G and compare transistor performance among fabs. The Intel gate structure creates some new problems for analysts.

Intel reports a gate length of 35 nm, which fits well with the 1.36- mA/µm drive current generated by its NFET. But the edge-to-edge dimension of the gate structure is closer to 45 nm if measured in a fashion similar to the standard used for polysilicon gates.



Click here for larger image

So what gives? The ratios of L_G, L_{elec} and source/drain extension lengths would be out of whack to produce such large saturation currents.

The answer appears related to our question about the location of the metal gate's edge. In the past, it was assumed that the entire width of the poly gate influenced carriers in the transistor channel. Because polysilicon is etched and replaced with a metal gate filling the trench in the gate-last process, the situation is less straightforward. The first material deposited into the gate trench is not metal for the gate but actually a barrier material, which means that the active portion of the gate is less than the traditional length measurement that would essentially run between the sidewall spacer on either side of the gate. However, because the barrier is quite thin, it cannot account for the gate measurement difference.

What appears to set the electrically active gate length is the bird's beak formed where the sidewall spacer meets the TIL. SI analysis concluded that this bird's beak is the result of TIL and high-k etches undercutting the polysilicon. Reoxidation of the polysilicon sidewall prior to silicon nitride spacer formation exacerbates the undercut. For the metal gate deposited into the trench, there is a thick, relatively low-k path toward the channel at this point that obviously could not electrically influence charge carriers in the region directly underneath the bird's beak.

The critical portion of the metal gate could also be the TIL itself. Because this layer is composed of the same work function metal as the gate-last layer, perhaps its edge defines the metal-gate length. Fortunately, the edge of the TIL layer approximately aligns with the bird's beak above it, so the choice of measurement point will not affect the value you get for L_G.

The punch line to all of this is that the gap between the gate trench edge and the electrically active edge of the work function metal (whether first or last) accounts for somewhere between 8 and 10 nm. And that appears to explain the difference between Intel's reported value for L_G and what the rest of us have been looking at.

Despite its cure for leakage power, adding hafnium creates headaches for the process integration engineer. Intel avoided hafnium's downsides (threshold voltage pinning and reduced carrier mobility) by creating a silicon oxide (or possibly oxynitride) bottom interface layer (BIL) between the silicon substrate and the HfO_2 layer. The BIL not only gets hafnium into the gate stack but gives the process engineer one more tuning knob. Because the gate dielectric's influence on the transistor channel and electrical performance is a function of the individual contributions of the various layers, threshold voltages can be controlled by varying the BIL thickness for different transistor applications.

Design-for-manufacturabilty
Process variability, and designing for it, is a hot topic as problems such as line-edge roughness and random dopant fluctuations become more problematic at 45 nm. This was addressed in Intel's second IEDM 2007 presentation, in which Kelin Kuhn discussed improving yield by process improvements as well as design changes. The SRAM cell illustrated Kuhn's point as she showed the evolution from 90-nm to 45-nm design. The "tall" cell layout used at 90 nm was replaced with a "wide" cell at 65 nm. The 65-nm cell design improved dimension control and variability by aligning the polysilicon in a single direction and removing the corners in the active area patterns. At 45 nm, Intel's process removed "dog bone" and "icicle" shapes by employing only square end caps. These uniform structures are also easier to fill reliably in the gate-last process.

Intel continues to use 193-nm dry lithography at 45 nm. Restricted design rules create "structured" gate layouts, as Kuhn mentioned in her discussion of the

SRAM cell. This DFM technique of uniform, regular arrangement of metal gates improves yields for the advanced HKMG technology without requiring investment in new immersion tooling. Creating strictly rectangular gate patterns did require an extra step, because double-patterning was used for the sacrificial polysilicon layer.

Many features of Intel's 65-nm process remain in evolved forms. "Third generation" strained silicon is used that is structurally similar to the embedded SiGe PMOS of Intel's 65-nm process. Nickel salicide is also used again at 45 nm. Intel employs dual damascene copper up to metal nine. A SiCN barrier with carbon-doped oxide creates the low-k interlevel dielectric integration scheme.

Don Scansen
(dons@semiconductor.com), semiconductor technology analyst at Semiconductor Insights, a CMP Technology company, holds a BSEE, MS and PhD from the University of Saskatchewan.